Extending Phone Prediction Models of Word Segmentation to a More Realistic Representation of Prosody

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for *graduation with research distinction* in Linguistics in the undergraduate colleges of the Ohio State University

By

John K. Pate

The Ohio State University

June 2009

Thesis Committee:

Chris Brew, Advisor

Mary Beckman

Eric Fosler-Lussier

Abstract

Laboratory results (e.g. Saffran et al. (1996)) have shown that infants can use statistical cues for word segmentation, and Christiansen et al. (1998) propose Simple Recurrent Networks (SRNs) as a model for this phenomenon that can integrate prosodic and segmental cues. I describe an augmentation of the corpus in Rytting (2007) with measurements of acoustic prosodic correlates, and perform a series of SRN phone prediction experiments on this enriched corpus. Although certain information-theoretic properties of this enhanced corpus suggest that prosodic correlates are predictive of word boundaries, two sets of experiments suggest that an SRN phone prediction task is an unsuitable basis for finding the strongest prosodic predictors of word boundaries. The first set explores manipulations on the inputs presented to the model, and the second set explores modifications of the model itself. I close by describing peculiarities of SRNs and the phone prediction task, and presenting desiderata of models for integrating acoustic correlates to prosody in word segmentation.

Acknowledgements

This thesis would not have been possible without Chris Brew, as he guided me to this topic as a starting point for understanding the role of prosody in the organization of linguistic objects. He has proven an invaluable source of encouragement and insight, and moreover has impressed upon me the importance of simplicity and clarity in both experimental design and presentation.

Many thanks also go to Anton Rytting in helping me understand his study, of which this is an extension, and for making available all the resources which he developed.

The role of Eric Fosler-Lussier has also been pivotal, as he introduced me to computational speech processing, and has been an eager source of insight into the biases and expectations inherent to different statistical learning algorithms.

Laura Wagner has helped me not only see the deeper questions relating to grammar representation and learning, but to appreciate the merits and deficits of conflicting approaches. Her willingness to forcefully advocate for viewpoints which I initially (and naïvely) dismiss has led me to a much more considered theoretical alignment on issues of language acquisition.

I also owe a debt to Cynthia Clopper instilling in me an appreciation for the staggering amount of variation that real language use exhibits, and showing me that increased variation can make tasks easier.

Vita

September 1986 ..... Born

June 2005 ..... Cincinnati Hills Christian Academy Highschool

Publications

John Pate and Detmar Meurers (2007). Refining Syntactic Categories Using Local Contexts – Experiments in Unlexicalized PCFG Parsing. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*. Bergen, Norway.

Field of Study

Major Field: Linguistics

Contents

List of Tables

List of Figures

Chapter 1

Introduction

One of the first tasks that faces a child in language acquisition is the division of utterances into smaller units like words and phrases. Several studies (e.g. Saffran et al. (1996); Thiessen and Saffran (2004); Saffran et al. (1999)) have established that children and adults successfully track transitional probabilities in acoustic signals to learn wordlike units. Since these laboratory studies typically employ artificial languages with computer-generated acoustics, however, they demonstrate this word-learning mechanism only in signals with acoustic and statistical cues that are strong, invariant, and inerrant.

Christiansen et al. (1998) used a connectionist learning paradigm to relax the assumptions of invariance and innerancy with regard to the statistical cues. A Simple Recurrent Network (SRN) was presented with a phone prediction task with a corpus of naturalistic Child Directed Speech, and easily outperformed intelligent baselines. Figure 1.1 presents the basic architecture of SRNs. They are essentially standard

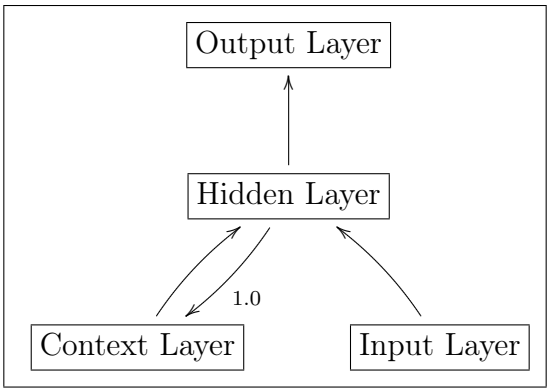


Figure 1.1: Schematic of Simple Recurrent Network

feedforward backpropagation networks with one hidden layer which have been augmented with a context layer. At each time step, the hidden layer is copied to the context layer (essentially freezing all weights from hidden nodes to context nodes at 1.0).

Figure 1.2 presents the structure of a basic phone prediction task. At each time step, the error of the network is calculated according to what the next phone is. Christiansen et al. (1998) include a special `ubm` phone that is inserted at each utterance boundary. Utterance boundaries are chosen because they have an obvious acoustic cue (a long period of silence) and typically always occur at word boundaries. The phone prediction task allows the network to exploit information about word boundaries relating to transitional probabilities, and reducing error on utterance boundaries allows networks to learn something about at least a subset of word boundaries.

Time step	1	2	3	4	5	6	7	8	9	10	11
Target	w	a	S	j	o	r	f	i	c	t	`ubm`
Input	`ubm`	w	a	S	j	o	r	f	i	c	t

Figure 1.2: Diagram of phone prediction task with "wash your feet"

SRNs have been used as a model of word segmentation for three primary reasons. First, as the hidden layer receives input from the context nodes along with the input nodes, the network can preserve sequential information. Because sequential information is preserved in a cycle, the "memory span" of the network is not limited to an *a priori* fixed window but is free to be learned and vary. Second, as the input nodes are simply real-valued numbers, it is a straightfoward process to provide the network with multiple sources of information that are only weakly predictive in order to see how readily these sources of information can be integrated. Finally, as SRNs exploit fairly general statistical cues without much innate knowledge, success at linguistic tasks with SRNs is often used to argue for a domain-general empiricist view of language acquisition.

Since a naturalistic corpus probably contains transitional probability distributions that reasonably approximate the messier distributions encountered by a child, Christiansen et al. (1998) established that statistical cues exist in what children hear with sufficient reliability and strength to be useful. However, as the phonemic transcriptions were drawn from a pronunciation dictionary, Christiansen et al. (1998) retain the assumptions of acoustic invariance.

Rytting (2007) relaxed the assumption of acoustic invariance using the same SRN phone prediction task. Rather than using the transcriptions from a pronunciation dictionary directly, the transcriptions were used to provide a forced phone alignment with actual audio files, and then the audio for each phone was used as the input into an Automatic Phone Classifier (APC). This resulted in, for each phone token,

a probability distribution over phones that respected acoustic variation. The phone symbols of Figure 1.2 are accordingly replaced by real-valued vectors. Upon relaxing the assumption of acoustic invariance, learners seeking to exploit cues from a phone prediction task to word boundaries in naturalistic CDS were found to be more severely impacted by highly variable input than previously thought.

Christiansen et al. (1998) also include a representation of lexical stress in their input. Since stress is often a cue to a syllable's location within a word, (e.g. stress-first in English c.f. Cutler and Carter (1987)), the possible utility of such acoustically-realized prosodic cues is worth investigating. However, the representation of stress used in Christiansen et al. (1998) is similarly problematic: it assumes that primary and secondary stress are always produced unambiguously and invariantly, and moreover draws the stress annotations directly from a dictionary, while the learner being modeled by hypothesis does not yet have access to such a dictionary.

The current study explores the utility of acoustically-realized cues to word boundaries in an SRN phone prediction task. Specifically, I augment the APC corpus employed in Rytting (2007) with eleven automatic measurements of known acoustic correlates to prosody, and see how useful these putative acoustic cues to prosody are for word segmentation in a phone prediction task. I report that these cues are difficult to use for word segmentation, and that advantages over segmental information alone do not appear except when the informativity of the input is highly constrained. Moreover, I find that these networks have a difficult time using the most powerful predictors of word boundaries.

Chapter 2

Materials

Rytting (2007) defined two sub-corpora containing utterances from four mothers of the Brent corpus from the Childes Database of child development corpora (Brent and Siskind, 2001; MacWhinney, 2000); a smaller corpus (`Brent60`) contained utterances whose phones had been classified with an accuracy of at least 60%, whereas a more variable corpus (`Brent33`) contained utterances whose phones had been classified with an accuracy of at least 33%. I use the larger and more variable `Brent33` corpus for this study. In this corpus, each phone is associated with a vector containing a probability distribution over 36 phones (the 34 phones of the MRC phoneset plus voiced and voiceless stop closures). Rytting (2007) contains full details about the corpus and handling of stop closures. I augment `Brent33` with prosodic correlates by appending to each phone vector eleven measurements obtained with a Praat script (Boersma and Weenink, 2008).

The measurements are selected to capture the four general acoustic properties that are understood to characterize prosodic objects: duration (one measurement), pitch (six measurements), voicing quality (three measurements), and volume (one measurement). Notably, Sluijter et al. (1997) show one of the voicing quality measures alone to be diagnostic of stress in Dutch (although Campbell and Beckman (1997) show that one measure to be insufficient in English). Thiessen and Saffran (2004) demonstrate that infants and adults perform word segmentation on the basis of these measures of pitch and voicing quality.

Specifically, the measurements gathered are:

- Duration
    1. Utterance-Normalized Duration: A z-score of the duration, in milliseconds, of the phone p using the mean and standard deviation of durations for the phones in the utterance containing p.
- Pitch
    1. Utterance F0 slope: The slope of a linear regression of mean phone F0 measurements from the same utterance against time.
    2. Utterance F0 linearity: The correlation coefficient of the linear regression of mean phone F0 against time.
    3. Utterance F0 linear error: The Root Mean Squared Error of the linear regression of mean phone F0 against time.

4. Utterance-Normalized F0: A z-score of the residual of the mean F0 of phone p from the linear regression normalized against the mean and standard deviation of residuals for the phones in the utterance containing p.
5. Utterance-Normalized F0 with zeros: This is identical to Utterance-Normalized F0 except phones with no F0 are assigned a score of 0.
6. Phone F0 slope: This is the difference in mean F0 from the final quartile of the phone duration minus the F0 from the first quartile of the phone duration divided by the total phone duration.

- Voicing quality
  1. H1-H2: This is the power of the first harmonic minus the power of the second harmonic. It is a measure of spectral tilt, which captures voicing quality.
  2. H2-H4: This is the power of the second harmonic minus the power of the fourth harmonic. It is another measure of spectral tilt.
  3. Gross tilt: This is the total energy above 500 Hz minus the total energy below 500 Hz, and is a third measure of spectral tilt.
- Volume
  1. Loudness: This is the total area under an excitation curve, which plots phons (a psycho-acoustically informed transform of intensity) against Barks (a psycho-acoustically informed transform of frequency). The transforms are necessary because the human ear is differentially sensitive to differences in raw acoustic measures at different ranges.

Some of the measures depend on the phone having a fundamental frequency and/or a harmonic series, but some phones (i.e. voiceless phones) do not have a fundamental frequency. In these cases, except where otherwise noted, the phones received the value of the most recent voiced phone; if there are no preceding voiced phones in the utterance, they received the value of the next voiced phone. The three measures of spectral tilt were selected on the basis of Kreiman et al. (2007), which compared 87 measures of spectral tilt in a Principle Components Analysis and concluded that these three plus a fourth more computationally expensive measure accounted for nearly all the variation. All measures of spectral tilt were gathered from spectra calculated over the entire phone duration.

All measures were linearly scaled to the interval $[0, 1]$ prior to presentation to the networks.

Chapter 3

Manipulating the Input

In this chapter, we maintain the SRN phone prediction model for word segmentation as formulated in Christiansen et al. (1998), and explore only its capacity to incorporate the more realistic representation of prosodic cues. This is achieved by presenting the model with different data sets while preserving model operation as in Christiansen et al. (1998) and Rytting (2007)[1].

3.1 Basic Experimental Set-up

The word segmentation task is modeled, following Christiansen et al. (1998) and Rytting (2007), as a corrollary of a phone prediction task with a special `ubm` symbol included in the phone set for "utterance boundary." The structure of our phone prediction task is similar to that of Figure 1.2 except, as in Rytting (2007), all symbols

[1] As a necessary consequence of using feature sets of different sizes, the number of nodes in the input layer will change. The number of nodes in the hidden and context layers will be adjusted accordingly to maintain a roughly constant number of model parameters

are replaced by vectors. Additionally, some networks receive prosodic measurements as inputs, while targets never contain prosodic measurements, and some networks do not receive phone information as inputs (receiving only prosodic measurements instead).

After training, network parameters are frozen and the SRN is run on an unseen test set. Following Christiansen et al. (1998) and Rytting (2007), the network is taken to predict a word boundary whenever the `ubm` node receives an activation that is above the average `ubm` activation over all positions. Following Rytting (2007), word boundaries posited between a stop closure and its burst are taken to have been posited before the stop closure, and all performance figures have been averaged over nine runs.

The training regime is similarly drawn from Christiansen et al. (1998). The SRN is trained to predict the next symbol for only one epoch to prevent convergence. If the SRN were to converge, few or no utterance-internal word boundaries would be predicted because utterance-internal boundaries are counted as an error during training. The hope is that the network will learn what word boundaries and utterance boundaries have in common (i.e. general phonotactics) before learning patterns that are specific to utterance boundaries. If network performance on word boundaries is higher than a baseline that assumes each utterance is one word, than I conclude that the most useful patterns (those that figure soonest in error minimization) for utterance boundaries are a reliable basis for generalizing to word boundaries.

Following previous work, I use a learning rate of 0.1 and momentum of 0.95, and

initialize weights randomly in $[-0.25, 0.25]$. Network performance is presented along with a baseline strategy which assumes that every utterance is one word for comparison.

3.2 Experiment I

Condition	Boundary			Lexical		
	P	R	F	P	R	F
Baseline	100	31.4	47.8	12.2	12.9	12.5
`seg`	47.0	70.9	56.6	16.7	30.1	21.5
`pros`	35.5	66.3	46.2	11.9	17.1	14.1
`pros-seg`	47.0	65.8	54.9	15.7	30.0	20.6

Table 3.1: Experiment I Boundary and Lexical Precision, Recall, and F-score for three Simple Recurrent Networks and a baseline.

Experiment I is a straightforward extension of previous work. I have three SRNs which differ in input and the size of the hidden and context layer. Our `seg` SRN receives segmental information in addition to the `ubm` node, and is a replication of one of the networks from Rytting (2007). The segmental information is the Automatically Phone Classified seventeen-feature input. Our `pros` network receives the acoustic measures described above in addition to the `ubm` node. Our `pros-seg` network receives as input both the seventeen-feature APC input, the acoustic measures of prosodic correlates, and the `ubm` node. As each network requires a different number of input nodes, we follow Rytting (2007) in manipulating the size of the hidden layer to keep the count of network parameters roughly constant. Table 3.2 describes our networks.

Name	In-Hidden-Out	# Parameters
`pros`	12–79–37	10,112
`seg`	18–77–37	10,164
`pros-seg`	29–73–37	10,147

Table 3.2: The three networks of Experiment I.

3.2.1 Results & Evaluation

Table 3.1 presents performance figures for each network and a baseline that assumes each utterance is one word. I provide Precision, Recall, and F-score for both word boundaries and word type. Precision is the percentage of guesses made by the network that were right; for example, `seg` obtains a boundary precision of 47.0%, which means that 47.0% of the boundaries it posited were right. Recall is the percentage of objects the network encountered which were successfully posited; for example, `seg` obtains a boundary recall of 70.9%, which means that 70.9% of real word boundaries were actually posited by the network. F-Score is the harmonic mean of Precision and Recall, and simply provides a canonical way to combine these two figures into one.

I follow Rytting (2007) in assuming the network successfully posits exactly one word boundary at every utterance boundary, regardless of the activation of the `ubm` node.

We see from Table 3.1 that all networks far outperform the baseline in boundary recall (the baseline by definition achieves perfect boundary precision). Moreover, all networks outperform the baseline in lexical recall by several points, and all but `pros` outperform the baseline in lexical precision by several points. As `pros` underperforms the baseline in lexical precision by only a very little amount, all networks outperform

the baseline in balanced F-score. Among our networks, I see that `seg` outperforms `pros` on every measure, often by substantial margins. Surprisingly, the `pros-seg` network marginally underperforms or, at best, matches the `seg` network on every measure.

3.2.2 Discussion

The disappointing performance of our `pros-seg` network suggests that either our prosodic cues do not contain very much information about word boundaries, or that the prosodically-grounded patterns for phone prediction do not generalize well to word boundary prediction. The predictive capacity of our prosodic measures for word boundaries can be quantified by calculating the entropy of a word boundary variable *wb*, which takes the value 1 for word-final phones, conditioned on our set of prosodic features *pros*:

$$\mathrm{H}(wb|pros) = \mathrm{H}(wb, pros) - \mathrm{H}(pros) \tag{3.1}$$

where

$$\mathrm{H}(wb, pros) = -\sum_{i,j} p(wb_i, pros_j) \log_2(p(wb_i, pros_j)) \tag{3.2}$$

and

$$\mathrm{H}(pros) = -\sum_{k} p(pros_k) \log_2(p(pros_k)) \tag{3.3}$$

As Equation 3.1 suggests, conditional entropy is simply the amount of information that is left over, in the best possible predictor, after we know the value of *pros*. p is defined in terms of relative frequencies. To produce a discrete approximation of our actual measures (which are continuous), I divide the range of each measure into

fifteen evenly-spaced bins and gather counts within those bins. Table 3.3 presents the overall entropy of word boundaries *wb* (which is an upper bound on the entropy of *wb* conditioned on anything), the entropy of *wb* conditioned on our feature sets, the overall entropy of utterance boundaries *ub*, and the entropy of *ub* conditioned on our feature sets for the entire corpus (training and test). The feature sets are *pros*, *seg*, and *pros-seg*, and are binned versions of the input to the similarly-named networks of Experiment I.

Variable	Feature set	Entropy
wb	$\emptyset$	0.892
	pros	0.161
	seg	0.383
	pros-seg	0.073
ub	$\emptyset$	0.382
	pros	0.037
	seg	0.076
	pros-seg	0.001

Table 3.3: Entropies.

We see from Table 3.3 that, while segmental features substantially reduce the entropy of word boundaries, prosodic features reduce the entropy of word boundaries even more, and both segmental and prosodic features together reduce the entropy of word boundaries even further. This suggests that our prosodic measures do in fact contain substantial information about word boundaries, and that the SRN is not generalizing well to word boundaries when performing the phone and utterance boundary prediction task.

Table 3.3 also presents the entropy of utterance boundaries conditioned on our feature sets. Notice that the entropy of utterance boundaries is substantially lower than

the entropy of word boundaries, meaning that utterance boundary prediction is substantially easier than word boundary prediction. Moreover, the entropy of utterance boundaries conditioned on prosodic correlates is very small, and the entropy of utterance boundaries conditioned on both prosodic correlates and segmental information is nearly zero. While prosodic correlates are very good predictors of word boundaries, they are superb predictors of utterance boundaries. Prosodic correlates may simply be such good predictors of utterance boundaries (or perhaps phrase boundaries) that patterns specific to utterance-boundaries emerge very early in training and dominate patterns that are common to word boundaries generally. This suggests that advantages to using prosodic correlates may become clearer under a more restricted training process.

3.3 Experiment II

Condition	Boundary			Lexical		
	P	R	F	P	R	F
Baseline	100	31.4	47.8	12.2	12.9	12.5
`seg`	43.8	76.8	55.8	14.6	22.3	17.6
`pros`	39.5	51.3	44.6	10.4	19.5	13.5
`pros-seg`	42.7	75.6	54.6	14.7	23.4	18.1

Table 3.4: Experiment II Boundary and Lexical Precision, Recall, and F-score for three Simple Recurrent Networks and a baseline.

In Experiment II, I restrict the networks to a small number of features, and select the particular feature set of a fixed size that minimizes the conditional entropy of utterance boundaries. Specifically, each of our `pros`, `seg`, and `pros-seg` networks receives 4 inputs. To determine which inputs to use for our `pros` and `seg` networks, I

calculate the entropy of utterance boundaries conditioned on every four-feature subset of the original feature sets, and pick the four-feature subset that results in the lowest conditional entropy. To select four features for our `pros-seg` network, I select the best set of two `pros` feaures and the best set of two `seg` features. The size of the hidden layer is manipulated to keep the number of parameters comparable to the networks of Experiment I and previous work. Table 3.5 presents information about the networks for Experiment II. Because the number of inputs is constant, however, architecture is exactly constant across networks in terms of both the number of nodes and the number of total parameters, meaning that this experiment better controls against details of SRN architecture.

Name	In-Hidden-Out	# Parameters
`pros`	4–82–37	10,086
`seg`	4–82–37	10,086
`pros-seg`	4–82–37	10,086

Table 3.5: The three networks of Experiment II.

3.3.1 Results

The networks are trained and evaluated in the same way and with the same measures as the networks from Experiment I.

I evaluate the networks with the same lexical and boundary Precision, Recall, and F-score for Experiment II are reported in Table 3.4. All networks predictably underperform their counterparts from Experiment I. `pros` achieves the lowest performance

on all measures. `seg-pros` matches or only slightly outperforms `seg` on lexical precision, but outperforms `seg` by a full point on lexical recall. Notably, although the `pros-seg` network lost the largest number of inputs, it exhibits the least degradation.

So, when the input contains only the strongest predictors of utterance boundaries, incorporating prosodic correlates leads to marginally better generalization to word boundaries in a phone prediction task. This tentatively confirms the suspicion that the multitude of prosodic correlates in Experiment I may have been prodding the networks to exploit patterns that were specific to utterance or phrase boundaries but exclusive of word boundaries.

3.4 Experiment III

In Experiment II, I found that, when the informativity of the inputs was restricted to feature sets that minimize the conditional entropy of utterance boundaries, prosodic and segmental cues together provided the best generalization to word boundaries. And throughout, networks which have received prosodic input have outperformed the baseline. So some prosodically-based cues to word boundaries can be discovered in a phone prediction task with SRNs.

However, in Experiment II, the improvement over the performance of segmentals alone was small, while Table 3.3 predicts a large gain from the incorporation of prosodic features. Moreover, Table 3.3 suggests that prosodic correlates alone are better predictors of word boundaries than segmental features alone, but the nets

which use prosodic correlates exclusively have only barely outperformed the baseline in Experiments I and II. It seems that our SRN phone prediction task is finding only the weakest prosodically-based predictors of word boundaries. Experiment III provides the networks with feature sets that contain the best patterns for predicting word boundaries, and sees if the SRN phone prediction task discovers these best patterns.

To see if a phoneme and utterance boundary prediction task will find the most useful prosodically-based patterns for predicting word boundaries, I rank four-member feature sets in the same fashion as was used for Experiment II except I pick input feature sets according to the conditional entropy of word boundaries (instead of utterance boundaries). If I tell the network which feature set contains the best patterns for predicting word boundaries, will a phone prediction task find those patterns? The networks have the same architecture here as in Experiment II.

3.4.1 Results

Table 3.6 presents the same boundary and lexical Precision, Recall, and F-score figures for our Experiment III nets that were presented for those in Experiments I and II.

We see in Table 3.6 further degradation in performance on all networks, suggesting that generalizing to using the best predictors for word boundaries is difficult without using the best predictors for utterance boundaries as well. This suggests that the most useful patterns for word boundaries are not apparent when looking at phone and utterance boundary prediction.

Condition	Boundary			Lexical		
	P	R	F	P	R	F
Baseline	100	31.4	47.8	12.2	12.9	12.5
`seg`	43.9	66.2	52.8	13.0	25.0	17.1
`pros`	38.5	54.4	45.1	10.1	12.5	11.2
`pros-seg`	40.4	78.9	53.4	13.7	20.1	16.3

Table 3.6: Experiment III Boundary and Lexical Precision, Recall, and F-score for three Simple Recurrent Networks and a baseline.

3.5 Discussion

Experiment I showed that this model as formulated in Christiansen et al. (1998) and Rytting (2007) struggles to use phonetically-motivated acoustic correlates to prosody for word segmentation. Specifically, model performance was best when presented with segmental cues only, and dropped when presented with both segmental and prosodic cues. As a subsequent information-theoretic analysis showed that the phonetically-motivated acoustic measures do contain information that should be for word-boundary classification, the failure of the model to incorporate the prosodic correlates into word segmentation is likely not an artifact of the particular feature set chosen.

Experiments II and III sought to acknowledge the relevance of prosodic cues to many levels of linguistic analysis. These experiments were motivated under the working hypothesis that these prosodic correlates worked in concert to push the model towards predicting some other sort of linguistic object, such as intonational phrase boundaries or major constituent boundaries. To counteract this superinformativity of our feature sets, the informativity of the input was dramatically reduced by throwing out most of the features. In Experiment II, the best four features for predicting utterance

boundaries were preserved, and the `pros-seg` network slightly outperformed the `seg` network. However, as the margin of improvement was very small, Experiment III looked at what would happen if the model were provided the best feature sets for predicting gold-standard word boundaries. Somewhat surprisingly, Experiment III produced the same pattern of results as Experiment I, with `pros` achieving the worst performance, `seg` achieving the best, and `pros-seg` achieving a slight degradation from `seg`.

These results suggest that the difficulties encountered by the model do not lie solely in how much information prosodic correlates contain about hidden linguistic phenomena. Instead, it may be the case that certain peculiarities of this model hinder the incorporation of prosodic cues. Chapter 4 details and evaluates some modifications to the model that were pursued in an attempt to address these peculiarities.

Chapter 4

Manipulating the Model

Chapter 3 presented three experiments which, together, suggest that the SRN phone prediction model for word segmentation of Christiansen et al. (1998) struggles to effectively utilize a realistic representation of prosodic cues, and that this difficulty does not lie entirely in the particular feature selection. This Chapter identifies one peculiarity of the model itself which may underlie its difficulty in incorporating prosodic cues, and presents some attempts to modify the model to overcome this peculiarity. Unfortunately, the modifications do not produce an improvement in performance, indicating that either more substantial changes to this model are necessary to employ a realistic representation of prosodic cues to word segmentation or other classes of models should be pursued.

4.1 Motivation

A striking peculiarity of this model, in contrast to other connectionist models, is that the networks are not allowed to train to convergence but are instead held to exactly one pass over the training set. As previously noted, training to convergence would lead the model to predict few or no utterance-internal boundaries since activation of the `ubm` node utterance-internally is counted as an error.

However, limiting exposure to training data has real consequences. Prosodic events are usually discussed with reference to syllables and groupings of syllables. As syllables consist of multiple speech sounds, and as one time step in this model corresponds to one speech sound, this model probably needs to learn dependencies over several time steps to appreciate prosodic events. SRNs learn sequence information through the cycle between the hidden layer and the context layer, which can, in principle, encode an arbitrary amount of history. Rodriguez (2003), however, show that Simple Recurrent Networks require substantial training to learn dependencies over five time steps using Wall Street Journal data. Accordingly, modifications of the model were explored which seek to allow multiple training passes without penalizing all utterance-internal word boundaries.

4.2 Basic Modification

All of these modifications involve changes to the supervision signal for the `ubm` node. At epoch 0, the supervision signal for the `ubm` node is 1 at utterance boundaries and

0 elsewhere, as in previous experiments. At epoch $n + 1$, however, the supervision signal for the `ubm` node is some monotonic mapping of the output of the `ubm` node in epoch n. As training proceeds, then, the network is rewarded for positing utterance-internal word boundaries, and the optimal solution no longer necessarily excludes all utterance-internal word boundaries. Intuitively, this can be thought of as an attempt to get the network to start trusting itself and to use what it has learned.

The modified model involves a monotonic mapping of the output of the `ubm` node so that the conservativeness with which the supervision signal is modified can be changed, and also so that the network can be biased to prefer activations of the `ubm` node near zero or near one.

Four different mappings are explored, and train for eight epochs each. Due to long training times, the networks were not trained to convergence. The figures in Rodriguez (2003), however, show that network performance with longer dependencies improves tonically with training length. The experiments presented here, then, do not present actual performance upon convergence, but do indicate whether each modification helps or hinders learning longer-distance dependencies. Eight epochs should be enough, as the total number of time steps of backpropagation in eight epochs of this data set $(756,680)$ is much larger than the total number of time steps of backpropagation Rodriguez (2003) report for learning dependencies over 7 time steps.

4.3 The Mappings

Three mappings to define the target `ubm` activation of time step i of epoch $n+1$ as a function of timestep i of epoch n:

- Identity: Completely trust the network's `ubm` activations:

$$\text{target}_{i,n+1}(\texttt{ubm}) = \text{activation}_{i,n}(\texttt{ubm})$$

- Two-sided Attenuation: Push targets away from a middle activation with a sigmoidal pattern. α can be manipulated to change the steepness of the slope. A plot of this function in $[0, 1]$ for $\alpha = 10$ appears in Figure 4.1:

$$\text{target}_{i,n+1}(\texttt{ubm}) = \frac{1}{1 + \exp(-\alpha \cdot (\text{activation}_{i,n}(\texttt{ubm})))}$$

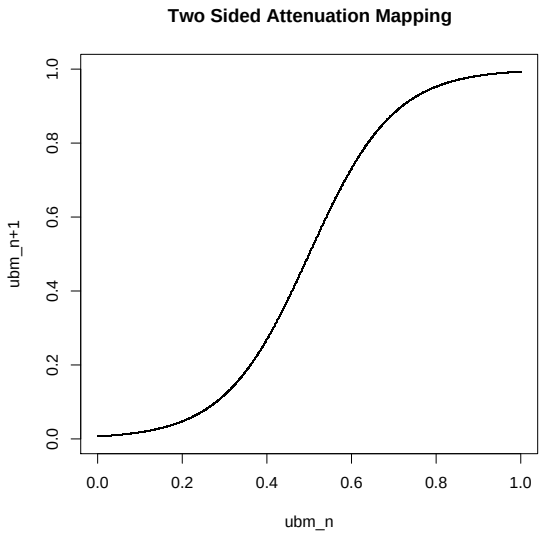
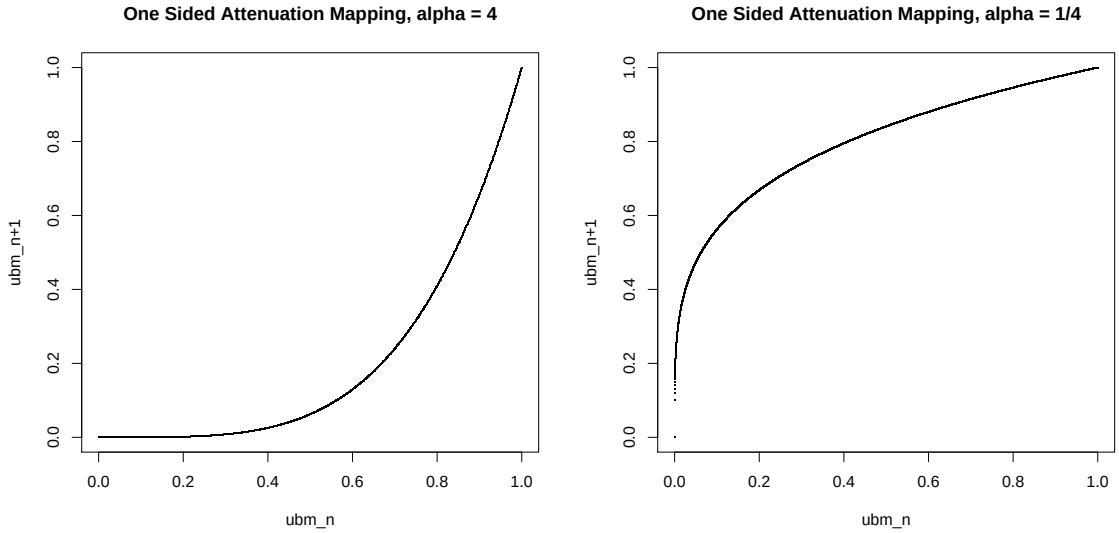


Figure 4.1: Two Sided Mapping

- One-sided Attenuation: Push targets away from either zero or one. When α is greater than one, targets are pushed away from one, and when α is less than one, targets are pushed away from zero. Intuitively, this tries to make the network require more evidence for either large or small activations, depending on α. Figure 4.2 plots this function for $\alpha \in \{\frac{1}{4}, 4\}$:

$$\text{target}_{i,n+1}(\texttt{ubm}) = \text{activation}_{i,n}(\texttt{ubm})^{\alpha}$$



(a) $\alpha = 4$ (b) $\alpha = \frac{1}{4}$

Figure 4.2: One Sided Mapping

4.4 Experimental Set-up

Many values of *alpha* for the attenuating mappings were tried, but little difference was found among them. Accordingly, I present results from one Two Sided Attenuation

network with $\alpha = 10$, a One Sided Attenuation network with $\alpha = 4$, a One Sided Attenuation network with $\alpha = \frac{1}{4}$, and an identity network. This section concerns itself with the ability of this modification of the model to incorporate segmental and prosodic cues, so figures are presented for the `pros-seg` data set only.

For comparison, I present figures for a baseline which assumes every utterance is one word, the `pros-seg` network trained for one epoch (copied from Chapter 3), the `pros-seg` network trained for two epochs (`pros-seg`.2), and the `pros-seg` network trained for eight epochs (`pros-seg`.8).

Training proceeded on the same 90%/10% training/test division with the same learning rate and momentum as before.

4.5 Results & Discussion

Condition	Boundary			Lexical		
	P	R	F	P	R	F
Baseline	100	31.4	47.8	12.2	12.9	12.5
`pros-seg`	47.0	65.8	54.9	15.7	30.0	20.6
`pros-seg`.2	45.9	62.5	52.9	15.0	29.2	19.8
`pros-seg`.8	39.0	76.4	51.6	15.3	23.8	18.6
Identity	39.0	77.4	51.9	14.9	23.0	18.0
Two-sided	38.9	78.1	52.0	15.7	23.5	18.8
One-sided ($\alpha = 4$)	39.2	78.7	52.3	15.0	22.3	17.9
One-sided ($\alpha = \frac{1}{4}$)	38.9	76.6	51.6	14.4	22.5	17.6

Table 4.1: Experiment I Boundary and Lexical Precision, Recall, and F-score Simple Recurrent Networks with and without a mapping modification.

Table 4.1 presents performance figures for these networks. Among the original `pros-seg` network, additional epochs of training lead to an improvement in only boundary recall, and this improvement is accompanied by a decrease in performance on all other figures.

The modified networks present a similar pattern, offsetting a clear jump in boundary recall with decreases in performance on the other measures. Strikingly, the `pros-seg` network (without the mapping modification) outperforms the networks with the mapping modification on most measures. Given the similar performance of other parametrizations of this technique (i.e. other values for α) to allow effective learning of long-distance dependencies, it seems that this SRN phone prediction model requires a more fundamental modifications before it is capable of using a realistic representation of prosodic cues.

Chapter 5

Conclusion & Future Work

I gathered measures of known acoustic correlates to prosody in an attempt to demonstrate the utility of putative prosodic cues to word boundaries, such as stress, in the absence of a dictionary explicitly providing those cues. I discovered that acoustic correlates to prosody cannot be straightforwardly combined with segmental information to increase performance, although prosodic correlates appear to be excellent information-theoretic predictors of word boundaries. The utility of prosodic cues for utterance boundaries over and above word boundaries, however, suggested that our networks were not emphasising those patterns that are useful for word boundaries generally.

To restrict the ability of our networks to learn several patterns in the hope of strengthening those that are useful for word boundaries, I severely limited the input to the network. This produced relatively little degradation in the performance of our `pros-seg` network, suggesting that the strongest patterns for phone and utterance boundary

prediction among prosodic and segmental cues in combination are also strong predictors of word boundaries.

Next, to see if the best patterns for word boundaries are easily generalizable from phone and utterance boundary prediction, I presented the networks with severely restricted input that consisted of the best predictors for gold-standard word boundaries. The networks performed poorly, suggesting that phone and utterance boundary prediction do not implicitly yield the most useful patterns for word boundaries.

Since the original model failed to employ prosodic correlates effectively even when provided only the best features for predicting word boundaries, I next explored modifications to the model itself. One likely cause of the failure of the model as originally formulated lay in its restriction of training to only one epoch. This restriction was imposed to prevent the network from learning only utterance-external word boundaries, and so I produced a modification of the model that was intended to allow longer training without penalizing the model for positing utterance-internal word boundaries. Intuitively, this was attempted by having the model trust itself about the appropriate activation of the `ubm` node from epoch to the next. Disappointingly, the modifications did not lead to successful exploitation of increased training.

These results suggest exploring other mechanisms for word boundary identification, and, given the limitations of our particular model, propose two closely-related desiderata. First, as already noted, I ought to explore models which more naturally learn long-range dependencies. The modification explored in Chapter 4 might be thought

of intuitively as an adaptation of Expectation-Maximization for SRNs, and the different sorts of mappings I explored can be thought of as adapting the notion of strong and weak priors to SRNs. Other models, such as the two-stage model of Goldwater (2007), effectively exploit these Bayesian notions directly and rigorously, and so may serve as a more suitable class of models for incorporating realistic prosodic cues.

Second, I ought to explore models which explicitly incorporate lexical information as it is learned and at the level of confidence with which it has been learned. Again, although SRNs have the capacity of representing full words, for the reasons given above, only very short words, if any, are likely to be so learned in practice. An explicit lexical representation could facilitate the incorporation of a particularly useful kind of long range dependency.

Bibliography

Boersma, P. and Weenink, D. (2008). Praat: Doing phonetics by computer (version 5.0.43) [computer program]. Retrieved December 9, 2008 from http://www.praat.org/.

Brent, M. R. and Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:31–44.

Campbell, N. and Beckman, M. (1997). Stress, prominence, and spectral tilt. *ESCA Workshop on Intonation: Theory, Models, and Applications*, pages 67–70.

Christiansen, M., Allen, J., and Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2/3):221–268.

Cutler, A. and Carter, D. (1987). The prosodic structure of initial syllables in english. In *ECST-1987*, pages 1207–1210.

Goldwater, S. (2007). *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University.

Kreiman, J., Gerratt, B. R., and nanzas Barroso, N. A. (2007). Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research*, 50:595–610.

MacWhinney, B. (2000). The CHILDES project: tools for analyzing talk.

Rodriguez, P. (2003). Comparing simple recurrent networks and *n*-grams in a large corpus. *Applied Intelligence*, 19:39–50.

Rytting, A. (2007). *Preserving Subsegmental Variation in Modeling Word Segmentation*. PhD thesis, The Ohio State University.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old-infants. *Science*, 274:1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70:27–52.

Sluijter, A. M. C., van Heuven, V. J., and Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1).

Thiessen, E. D. and Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception & Psychophysics*, 66(5):779–791.